1. (Previously presented) A computer-implemented method of identifying table data in a document comprising the steps of:

receiving a page description language representation of the document for providing a list of words in the document and position information for the words; and

automatically identifying table data in the document based on the page description language representation of the document and at least one table identifying feature, wherein the identifying step includes,

dividing the document into one or more pages;

dividing each page into a plurality of lines;

for each line, clustering the words of the line into one or more word clusters, wherein each cluster includes one or more words, each cluster having a horizontal beginning point, horizontal midpoint, and horizontal end point;

for clusters in the plurality of lines, comparing alignment of the horizontal beginning point, horizontal midpoint, and horizontal end point of clusters between lines, wherein a cluster in a first line is aligned with a cluster in a previous line if at least one of the horizontal beginning point, horizontal midpoint, and horizontal end point of the cluster in the first line is aligned with at least one of the horizontal beginning point, horizontal midpoint, and horizontal end point of the cluster in the previous line; and

identifying a line as being part of a table in response to more than one cluster of the line being aligned with clusters of previous lines identified as part of the table; and

outputting data descriptive of the lines of the table.

2. (cancelled)

3. (previously presented) The method of Claim 1 wherein the step of automatically identifying table data in the document based on the number of word clusters of each line and the alignment of the word clusters between lines further comprises:

using the word clusters to generate column position information, wherein the column information includes for each column a horizontal beginning point, horizontal midpoint, and horizontal end point; and

updating the column position information by performing a union operation between the column position information of a previous line and the column position information of a current line.

4. (cancelled)

5. (previously presented) The method of Claim 1 wherein receiving a page description language representation of the document for providing a list of words in the document and position information for the words includes receiving a PDF representation of the document, and wherein converting the table data encompassed by each table bounding box to a markup language representation includes converting the table data encompassed by each table bounding box to a HTML representation.

6. (cancelled)

7. (Previously presented) A computer-readable medium having stored thereon sequences of instructions, said sequences of instructions including instructions which, when executed by a processor, cause said processor to perform the steps of:

receiving a page description language representation of a document for providing a list of words in the document and position information for the words; and

automatically identifying table data in the document based on the page description language representation of the document and at least one table identifying feature, wherein identifying includes,

dividing the document into one or more pages;

dividing each page into a plurality of lines;

for each line, clustering the words of the line into one or more word clusters, wherein each cluster includes one or more words, each cluster having a horizontal beginning point, horizontal midpoint, and horizontal end point; and

for clusters in the plurality of lines, comparing alignment of the horizontal beginning point, horizontal midpoint, and horizontal end point of clusters between lines, wherein a cluster in a first line is aligned with a cluster in a previous line if at least one of the horizontal beginning point, horizontal midpoint, and horizontal end point of the cluster in the first line is aligned with at least one of the horizontal beginning point, horizontal midpoint, and horizontal end point of the cluster in the previous line; and

identifying a line as being part of a table in response to more than one cluster of the line being aligned with clusters of previous lines identified as part of the table; and

outputting data descriptive of the lines of the table.

8. (cancelled)

9. (previously presented) The computer-readable medium of Claim 7 further containing instructions which, when executed by said processor, would cause said processor to perform the steps of:

using the word clusters to generate column position information, wherein the column information includes for each column a horizontal beginning point, horizontal midpoint, and horizontal end point; and

updating the column position information by performing a union operation between the column position information of a previous line and the column position information of a current line.

10. (cancelled)

11. (cancelled)

12. (Previously presented) A document processing system comprising:

a processor for executing programs; and

a table identification program for receiving a page description language representation of a document, the page description language representation providing a

list of words in the document and position information for the words, and for automatically identifying table data in the document based on the page description representation of the document and at least one table identifying feature, wherein the identification program is configured to,

divide the document into one or more pages;

divide each page into a plurality of lines;

for each line, cluster the words of the line into one or more word clusters, wherein each cluster includes one or more words, each cluster having a horizontal beginning point, horizontal midpoint, and horizontal end point;

for clusters in the plurality of lines, compare alignment of the horizontal beginning point, horizontal midpoint, and horizontal end point of clusters between lines, wherein a cluster in a first line is aligned with a cluster in a previous line if at least one of the horizontal beginning point, horizontal midpoint, and horizontal end point of the cluster in the first line is aligned with at least one of the horizontal beginning point, horizontal midpoint, and horizontal end point of the cluster in the previous line; and

identify a line as being part of a table in response to more than one cluster of the line being aligned with clusters of previous lines identified as part of the table; and

output data descriptive of the lines of the table.

13. (cancelled)

14. (cancelled)

15. (previously presented) The document processing system of claim 12 wherein the table identification program further comprises:

a conversion module coupled to the bounding box generation module for receiving the table bounding box for each table in the document, and for converting the words encompassed by the table bounding box into a markup language representation that maintains the table structure of each table.

16. (previously presented) The method of claim 1 wherein the step of automatically identifying table data in the document based on the page description language representation of the document and at least one table identifying feature further comprises:

automatically identifying table data in the document based on one or more table headings.

17. (previously presented) The method of claim 1 wherein the step of automatically identifying table data in the document based on the page description language representation of the document and at least one table identifying feature further comprises:

automatically identifying table data in the document based on one or more horizontal lines and vertical lines that separate rows or columns of the table.

18. (previously presented) The method of claim 1, wherein the step of automatically identifying table data in the document based on the number of word clusters for each line and the alignment of the word clusters comprises:

determining whether the number of word clusters in a line is greater than a threshold value; and

classifying the word clusters in the line as a row of a table in response to the number of word clusters in a line being greater than the threshold value.

19. (previously presented) The computer-readable medium of claim 7, wherein the instructions for automatically identifying table data in the document based on the number of word clusters for each line and the alignment of the word clusters include instructions that when executed by a processor cause the processor to perform the steps further comprising:

determining whether the number of word clusters in a line is greater than a threshold value; and

classifying the word clusters in the line as a row of a table in response to the number of word clusters in a line being greater than the threshold value.

20. (previously presented) The document processing system of claim 12, wherein the table identification program is further configured to:

determine whether the number of word clusters in a line is greater than a threshold value; and

classify the word clusters in the line as a row of a table in response to the number of word clusters in a line being greater than the threshold value.